Modeling document features for expert finding

Modeling Document Features for Expert Finding

Jianhan Zhu
University College London
Adastral Park Campus
Ipswich, IP5 3RE, UK
j.zhu@adastral.ucl.ac.uk

Dawei Song, Stefan Rüger
The Open University
Milton Keynes, MK7 6AA, UK
{d.song, s.rueger}
@open.ac.uk

Xiangji Huang
School of Information Technology
York University
Toronto, Canada
jhuang@yorku.ca

ABSTRACT
We argue that expert finding is sensitive to multiple document features in an organization, and therefore, can benefit from the incorporation of these document features. We propose a unified language model, which integrates multiple document features, namely, multiple levels of associations, PageRank, indegree, internal document structure, and URL length. Our experiments on two TREC Enterprise Track collections, i.e., the W3C and CSIRO datasets, demonstrate that the natures of the two organizational intranets and two types of expert finding tasks, i.e., *key contact* finding for CSIRO and *knowledgeable person* finding for W3C, influence the effectiveness of different document features. Our work provides insights into which document features work for certain types of expert finding tasks, and helps design expert finding strategies that are effective for different scenarios.

Categories and Subject Descriptors
H.3 [**Information Storage and Retrieval**]: H3.1 Content analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms
Algorithms, Experimentation, Measurement, Performance

Keywords
Expert finding, language models, enterprise search

1. INTRODUCTION AND MOTIVATION
Expert finding is a key task in enterprise search and has recently attracted lots of attention as evidenced by organization of expert search tasks in the Text REtrieval Conference (TREC) [1, 5, 7].

A prominent language modeling approach has been proposed by Balog et al [2] for expert finding. Petkova and Croft [6] further improved the approach by proposing a proximity-based document representation for incorporating sequential information.

However, these language modeling approaches have not sufficiently considered the effect of document features. In this paper, we propose a unified language modeling approach for taking into account the following document features in expert finding.

1. Internal document structure. For example, a person's name occurs in the author, content, reference, or acknowledgement section of a paper.

2. Document URLs. We consider URL length.

3. PageRank and indegree were shown to be effective for document retrieval [4]. We will study their effect in expert finding.

4. Anchor texts were shown to help Web search [3]. We will study their effect in expert finding.

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10.

5. Multiple levels of associations. Proximity between occurrences of an expert and topic terms is an indicator of the expert's relevance to the topic. We will study the effect of multiple levels of associations, and their integration with internal document structure in expert finding.

2. MODELING DOCUMENT FEATURES
Our models are instances of document-centric generative language modeling approaches to rank experts. Formally, given a set D of documents d, a query topic q, and a set C of candidates, the aim is to rank candidates based on the probability $p(c|q)=p(c,q)/p(q)$.

We adopt a document-centric generative language modeling approach by representing the joint $p(c,q)$ as a weighted average of the document models.

$$p(c,q) = \sum_{d \in D} p(c,q \mid d)\, p(d) \tag{1}$$

Most previous approaches ignore the prior $p(d)$ by assuming that it is uniform for all documents. However, we argue that the estimation of $p(d)$ based on multiple features of d such as its URL length, indegree, and PageRank etc. can influence the performance of expert finding. Assuming PageRank, indegree, and URL length are independent features, we estimate $p(d)$ as

$$p(d) \propto f_{PR}(d) f_{URL}(d) \quad \text{or} \quad p(d) \propto f_{indegree}(d) f_{URL}(d), \tag{2}$$

where $f_{PR}(d)$, $f_{URL}(d)$, and $f_{indegree}(d)$ are the *sigm* transformation functions proposed by Craswell et al [4] for PageRank, URL length, and indegree, respectively. We have

$$f_{Feature}(d) \propto w \frac{Feature(d)^a}{k^a + Feature(d)^a}, \tag{3}$$

where *Feature* is the PageRank, URL length, or Indegree of d, and w, a and k are parameters for the *sigm* transformation.

Now we focus on estimating $p(c,q|d)$ in Eq. 1.

$$p(c,q \mid d) = p(c \mid q,d)\, p(q \mid d) \tag{4}$$

$p(q|d)$ is estimated by inferring a document language model θ_d for each document d such that

$$p(q \mid \theta_d) = \prod_{t \in q} p(t \mid \theta_d)^{n(t,q)}, \tag{5}$$

where t is a query term and $n(t,q)$ is its frequency in q. We propose using a mixture of components to represent each document. By focusing on the effect of anchor text, we get

$$p(t \mid \theta_d) = (1-\lambda_c)(\lambda_t p(t \mid d_{text}) + \lambda_a p(t \mid d_{anchor})) + \lambda_c p(t), \tag{6}$$

where the document content part is weighted by $(1-\lambda_c)\lambda_t$, anchor text part is weighted by $(1-\lambda_c)\lambda_a$, $\lambda_t+\lambda_a=1.0$, and $p(t)$ is the maximum likelihood estimate of the term t given the background model, weighted with λ_c.

$p(c|q,d)$ in Eq. 4 denotes a co-occurrence model, which we construct as a linear interpolation of $p(c|d,q)$ and the background model $p(c)$ to ensure there are no zero probabilities as

$$p(c \mid \theta_d, \theta_q) = (1-\mu)\, p(c \mid d,q) + \mu p(c), \tag{7}$$

where $p(c)$ is the probability of candidate c. We estimate $p(c)$ as

$$p(c) = \frac{1}{df_c} \sum_{d' \in D} \frac{f(c,d')}{\sum_{c' \in C} f(c',d')}, \quad (8)$$

where $f(c,d')$ is the frequency of candidate c in document d' and df_c is the document frequency of c.

We use a Dirichlet prior for the smoothing parameter μ

$$\mu = \frac{\kappa}{\sum_{c' \in C} f(c',d') + \kappa}, \quad (9)$$

where κ is the average term frequency of all candidates in the corpus.

We use a multiple window based approach in estimating $p(c|d,q)$. We assume that small windows often lead to more probable associations, and large windows result in nosier associations, and weight smaller windows higher than larger ones.

Given a list W consisting of N windows $\{w_i\}$ $(i=1, ..., N)$ of different sizes, we estimate $p(c|d,q)$ as

$$p(c \mid d,q) = \sum_{w} p(w) p(c \mid d,q,w), \quad (10)$$

where $p(w)$ is the probability for each of the window-based co-occurrence models.

Based on the nature of the section where c is mentioned in a document, we combine the internal document structure information with the window-based co-occurrence model. Given a number of text windows where c co-occurs with q as $\{w_i\}$, we estimate $p(c|q,d,w)$ as follows

$$p(c \mid d,q,w) = \sum_{w_i} \frac{f(c,d,q,w_i)}{\sum_{c' \in C} f(c',d,q,w_i)}, \quad (11)$$

where $\sum_{c' \in C} f(c',d,q,w_i)$ is the total frequency of candidates in w_i.

Given a number of occurrences of c in w_i as $\{c_j\}$, $f(c,d,q,w_i)$ is estimated by combining internal document structure as

$$f(c,d,q,w_i) = \sum_{c_j} \delta(Section(c_j)), \quad (12)$$

where $\delta(Section(c_j))$ is a weighting function given to the section where c_j occurs, e.g., higher weight to occurrences of c in the author section of a technical paper.

3. EXPERIMENTAL FINDINGS

Based on our experiments the TREC2006 expert search collection on the W3C dataset and TREC2007 expert Search collection on the CSIRO dataset, we have the following findings.

Our expert finding approach achieves superior results in terms of our best MAPs on the two TREC datasets that are both better than previous language model based approaches [1] and those of the best automatic two-stage model runs in the TREC2006 and TREC2007 expert search tasks [1,7], respectively, even without using techniques such as query processing and query expansion.

In order to achieve good MAP, the window size used for association discovery should be sufficiently large, e.g., above 100 terms, for both collections.

Expert finding on the CSIRO dataset is a *key contact* search where very few experts per topic are defined, while expert finding on the W3C dataset is a *knowledgeable person* search where dozens of experts per topic are typical. Based on this difference, medium sized window should be used for *key contact* search since these *key contact* associations with the topic are more focused within medium range, and large windows introduce more noise than useful information. Associations of *knowledgeable person* and a topic tend to distribute more evenly across multiple windows, therefore, large window sizes should be used.

Anchor texts are more useful in key contact search since key contacts often appear in authoritative documents which attract inlinks, therefore, anchor texts. We found that an increased weight of anchor text leads to better performance than a pure document content based approach for large window sizes. However, anchor texts are less effective for knowledgeable person search since many experts may not appear in authoritative documents.

URL length is less effective than PageRank and indegree, which is also the case in document retrieval [4]. Due to the strong correlations between PageRank/indegree and document authority, they are both effective for key contact search, but less effective for all knowledgeable person search.

The rich internal structures of documents in the W3C dataset help improve expert finding with statistical significance, signifying its importance in expert finding on structurally rich datasets. Internal structures and indegree are complementary in expert finding since they describe different aspects of documents.

Window combination is effective for expert finding on the W3C dataset showing the wide distribution of expertise associations on different ranges, while less effective on the CSIRO dataset due to the concentration of expertise associations in small and medium ranges.

4. CONCLUSIONS AND FUTURE WORK

In order to develop generic expert finding approaches applicable to different scenarios, we have demonstrated that it is important and beneficial to study the effect of multiple document features. We proposed a novel approach of integrating document features in a language model for expert finding, and carried out a systematic investigation of the effects of document features in expert finding on two TREC test collections.

In future work, we plan to study the effect of query expansion and its relationships with document features, and effective multiple window combination method. In integrating PageRank, URL length, and indegree, we will investigate different transformation functions and explore the effect of parameters in the transformation functions in expert finding.

5. REFERENCES

[1] Bailey, P., Craswell, N., de Vries, A. P., Soboroff, I. (2008) Overview of the TREC 2007 Enterprise Track. In TREC 2007.

[2] Balog, K., Azzopardi, L. and de Rijke, M. (2006) Formal models for expert finding in enterprise corpora. In SIGIR 2006: 43-50.

[3] Craswell, N., and Hawking, D. (2005) Overview of the TREC-2004 Web Track. In TREC 2004.

[4] Craswell, N., Robertson, S.E., Zaragoza, H., and Taylor, M. J. (2005) Relevance weighting for query independent evidence. In SIGIR: 416-423.

[5] Craswell, N., de Vries, A.P., Soboroff, I. (2006) Overview of the TREC-2005 Enterprise Track. In TREC 2005.

[6] Petkova, D., and Croft, W. B. (2007) Proximity-based document representation for named entity retrieval. In CIKM: 731-740.

[7] Soboroff, I., de Vries, A.P. and Craswell, N. (2007) Overview of the TREC 2006 Enterprise Track. In TREC 2006.